PROYECTO IA - ENTREGA 2

Integrantes:

Angela Maria Gómez Londoño CC: 1026159266
Manuela Zapata Cifuentes CC: 1040326871
Felipe Barrera Hincapié CC: 1007223451

Descripción del progreso alcanzado

Preprocesamiento del dataset.

El dataset "Respiratory sound database" está compuesto de los siguientes archivos:

- 920 archivos de sonido .wav
- 920 archivos .txt
- Un archivo de texto con el diagnóstico de cada paciente
- Un archivo de texto que explica el formato de nomenclatura de los archivos
- Un archivo de texto con una lista de 91 nombres (filename_differences.txt)
- Un archivo de texto con información demográfica de cada paciente

Antes de realizar modelos predictivos es necesario el preprocesamiento de los datos. Preprocesar los datos se trata de cualquier tipo de procesamiento que se realiza con los datos brutos para transformarlos en datos que tengan formatos que sean más fáciles de utilizar.
Para el primer paso en el preprocesamiento se tomaron los datos presentes en el archivo "`patient_diagnosis.csv`" que está compuesto de 2 columnas. La primera es el ID de cada paciente, y la segunda es la enfermedad que tienen. En la columna de enfermedad se encuentran los posibles siguientes valores:

- COPD: Chronic Obstructive Pulmonary Disease
- LRTI: Lower Respiratory Tract Infection
- URTI: Upper Respiratory Tract Infection
- Asthma
- Healthy

	pid	Enfermedad
0	101	URTI
1	102	Healthy
2	103	Asthma
3	104	COPD
4	105	URTI

Figura 1: Composición del archivo "patient_diagnosis.csv"

Los 920 archivos txt según nos dice la descripción del dataset poseen 4 columnas que significan lo siguiente:

- Comienzo de los ciclos respiratorios
- Fin del ciclo respiratorio
- Presencia/ausencia de crepitancias (presencia=1, ausencia=0)
- Presencia/ausencia de sibilancias (presencia=1, ausencia=0)

0	3.250	6.636	0	0
1	6.636	11.179	0	1
2	11.179	14.250	0	1
3	14.250	16.993	0	1
4	16.993	19.979	0	0

Figura 2: composición de los archivos .txt

Además de esto, los nombres de los archivos nos brindan 5 informaciones diferentes separadas por un guión bajo. La información que nos brindan es: número del paciente, índice de grabación, ubicación, modo de adquisición y equipo de grabación. Para tener todos estos datos en un solo dataframe, usamos el método merge para juntarlos.

	inicio	fin	crepitancias	sibilancias	pid	modo	nombre	Enfermedad
0	1.862	5.718	0	1	160	mc	160_1b3_Al_mc_AKGC417L	COPD
1	5.718	9.725	1	1	160	mc	160_1b3_Al_mc_AKGC417L	COPD
2	9.725	13.614	0	1	160	mc	160_1b3_Al_mc_AKGC417L	COPD
3	13.614	17.671	0	1	160	mc	160_1b3_Al_mc_AKGC417L	COPD
4	17.671	19.541	0	0	160	mc	160_1b3_Al_mc_AKGC417L	COPD

Figura 3: dataframe final de archivos de texto.

Para el preprocesamiento de los audios se usaron los 920 archivos txt, los cuales, como ya se mencionó poseen 4 columnas, siendo en este caso las de mayor interés, el comienzo del ciclo respiratorio (“inicio”) y el fin del ciclo respiratorio (“fin”), esto debido a que se desea clasificar la frecuencia respiratoria como alta, baja o normal, según sea el caso. Por tal motivo, se creó un dataframe que tuviera estas 2 columnas, y adicional a estas, la longitud del ciclo respiratorio. Dicha longitud se obtuvo al restar el tiempo en que inició y finalizó el ciclo.

	inicio	fin	Longitud de ciclo	filname
500	9.365636	11.174727	1.809091	101
501	8.639333	10.300000	1.660667	101
512	9.807231	11.322615	1.515385	102
507	8.778667	12.047833	3.269167	103
510	11.984171	13.811599	1.827429	104
...	...	...	...	...
495	9.980571	12.831571	2.851000	224
497	9.587714	11.012214	1.424500	225
494	9.648818	11.461818	1.813000	226
496	8.444400	10.438600	1.994200	226
498	8.942800	10.927100	1.984300	226

920 rows × 4 columns

Figura 4: dataframe ciclos respiratorios.

Para el preprocesamiento de archivos de datos demográficos, se generó un dataset con los datos contenidos en el archivo "demographic_info.txt" de los cuales se extrajeron datos como la edad, el género y el índice de masa corporal. Estos datos se ingresaron a un dataframe como se muestra a continuación

	pid	Edad	Genero	BMI (kg/m2)	Enfermedad
0	101	3.00	F	0.00	URTI
1	102	0.75	F	0.00	Healthy
2	103	70.00	F	33.00	Asthma
3	104	70.00	F	28.47	COPD
4	105	7.00	F	0.00	URTI
5	106	73.00	F	21.00	COPD
6	107	75.00	F	33.70	COPD
7	108	3.00	M	0.00	LRTI
8	109	84.00	F	33.53	COPD
9	110	75.00	M	25.21	COPD

Figura 4: dataframe datos demográficos.